This document is the accepted manuscript version of the following article:
Podgorski, J., & Berg, M. (2020). Global threat of arsenic in groundwater. Science, 368(6493), 845-850. https://doi.org/10.1126/science.aba1510

Global threat of arsenic in groundwater

Joel Podgorski[1,2*] *and Michael Berg*[1,3*]

1. Eawag, Swiss Federal Institute of Aquatic Science and Technology, Department Water Resources and Drinking Water, 8600 Dübendorf, Switzerland

2. University of Manchester, Department of Earth and Environmental Sciences, Manchester, United Kingdom

3. UNESCO Chair on Groundwater Arsenic within the 2030 Agenda for Sustainable Development and School of Civil Engineering and Surveying, University of Southern Queensland, 4350 QLD, Australia

**Corresponding authors. Email: joel.podgorski@eawag.ch, michael.berg@eawag.ch.*

Revision for "*Science*"

Abstract:

Naturally occurring arsenic in groundwater affects millions of people worldwide. We created a global prediction map of groundwater arsenic exceeding 10 µg/L using a random forest machine learning model based on eleven geospatial environmental parameters and over 50,000 aggregated data points of measured groundwater arsenic concentration. Our global prediction map includes known arsenic-affected areas and previously undocumented areas of concern. Combining the global arsenic prediction model with household groundwater-usage statistics, we estimate that 94-220 million people are potentially exposed to high arsenic concentrations in groundwater, the vast majority being in Asia (94%). Since groundwater is increasingly utilized to support growing populations and buffer against water scarcity due to changing climate, this work is important to raise awareness, identify areas for safe wells, and help prioritize testing.

The natural, or geogenic, occurrence of arsenic in groundwater is a global problem with wide-ranging health effects for humans and wildlife. Being toxic and not serving any beneficial metabolic function, inorganic arsenic (the species present in groundwater) can lead to disorders of the skin and vascular and nervous systems as well as cancer(*1, 2*). The major source of inorganic arsenic in the diet is through arsenic-contaminated water, although ingestion through food, particularly rice, represents another important route of exposure(*3*). As a consequence, the World Health Organization (WHO) has set a guideline concentration of 10 µg/L in drinking water(*4*).

At least trace amounts of arsenic occur in virtually all rocks and sediments around the world(*5*). However, in most of the large-scale cases of geogenic arsenic contamination in groundwater, arsenic accumulates in aquifers composed of recently deposited alluvial sediments. Under anoxic conditions, arsenic is released from the microbial and/or chemical reductive dissolution of arsenic-bearing iron(III) minerals in the aquifer sediments(*6-9*). Also under oxidizing high-pH conditions, arsenic can desorb from iron and aluminum hydroxides(*10*). Furthermore, aquifers in flat-lying sedimentary sequences generally have a small hydraulic gradient, causing groundwater to flow slowly. This longer groundwater residence time allows dissolved arsenic to accumulate and its concentration to increase. Other processes responsible for arsenic release into groundwater include oxidation of arsenic-bearing sulfide minerals as well as release from arsenic-enriched geothermal deposits.

The fact that arsenic is generally not included in the standard suite of tested water quality parameters(*11*) and is not detected by the human senses mean that arsenic is regularly being discovered in new areas. Since one of the greatest occurrences of geogenic groundwater arsenic was discovered in 1993 in the Bengal Delta (*5, 12, 13*), high arsenic concentrations have been detected all around the world, with hotspots including Argentina(*14-17*), Cambodia(*18, 19*), China(*20-22*), India(*23-25*), Mexico(*26, 27*), Pakistan(*28, 29*), the USA(*30, 31*) and Vietnam(*32, 33*).

In order to help identify areas likely to contain high concentrations of arsenic in groundwater, several researchers have used statistical learning methods to create arsenic prediction maps based on available datasets of measured arsenic concentrations and relevant geospatial parameters. Previous studies have focused on Burkina Faso(*34*), China(*21, 35*), South Asia(*29, 36*), Southeast Asia(*37*), the USA(*31, 38, 39*) and the Red River Delta in Vietnam(*33*) as well as sedimentary basins around the world (*40*). The predictor variables used in these studies generally include various climate and soil parameters, geology and topography (Table S3).

Taking advantage of the increasing availability of high-resolution datasets of relevant environmental parameters, we use statistical learning to model what to our knowledge is the most spatially extensive compilation of arsenic measurements in groundwater assembled, which makes a global model possible. In order to focus on health risks, we consider the probability of arsenic in groundwater exceeding the WHO guideline. For this we have chosen the random forest method, which our preliminary tests showed to be highly effective in addressing this classification problem. We use the resulting model to produce the most

accurate and detailed global prediction map to date of geogenic groundwater arsenic, which can be used to help identify previously unknown areas of arsenic contamination as well as more clearly delineate the scope of this global problem and considerably increase awareness.

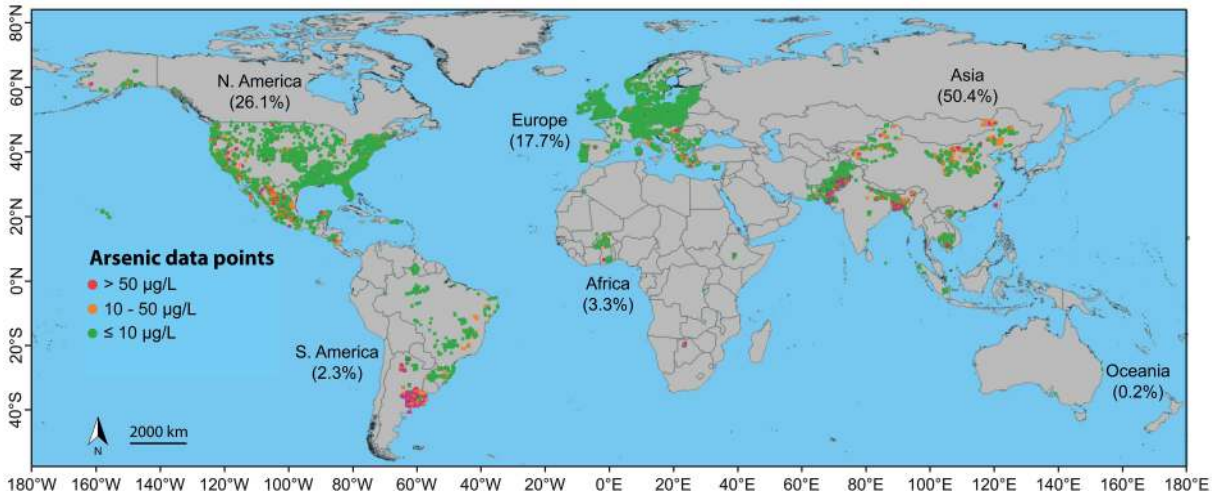


Figure 1. Arsenic concentrations excluding those known to originate from a depth greater than 100 m. Values are from the sources listed in Table S1. The geographical distribution of data is indicated by continent.

RESULTS

Random forest modeling

We aggregated data from nearly 80 studies of arsenic in groundwater (see Table S1 for references and statistics) into a single dataset (n>200,000). Averaging into 1 km^2 pixels resulted in more than 55,000 arsenic data points for use in modeling based on groundwater samples not known to originate from greater than 100 m depth (Figure 1).

In order to create the simplest and most accurate model, an initial set of 52 potentially relevant environmental predictor variables was iteratively reduced in consideration of their relative importance and impact on the accuracy of a succession of random forest models. The final selection of 11 predictor variables (Table S2) includes several soil parameters (topsoil clay, subsoil sand, pH and fluvisols), all of the climate variables (precipitation, actual and potential evapotranspiration and combinations thereof as well as temperature) and the topographic wetness index. In contrast, none of the geology variables proved to be statistically important. This is not to imply that geology does not play a role in geogenic arsenic accumulation, but rather that the particular geology variables tested were not as relevant as the other variables. This may be due to the coarse nature of the geological maps, which are standardized for the entire world. Although the number of predictor variables was

reduced by nearly 80%, both the Area Under the Curve (AUC, 0.89) and Cohen's kappa statistic (0.55) remained unchanged.

The final random forest model was created based on the compiled global dataset of high and low arsenic concentrations along with the 11 predictor variables. The standard number of variables to be made available at each branch of each tree is between three and four (see Methods). Since our tests showed the value of three performing better than four and higher values (though error/performance rates varied only within ~1%), we set this parameter to three. The global map produced from this model is displayed in Figure 2a along with more detailed views of the more populated affected continental regions (Figure 2b-f). It indicates the probability of the concentration of arsenic in groundwater in a given 1 km^2 cell exceeding 10 µg/L. The uncertainty of the model is inherent in the probabilities themselves, since they are simply the average of the votes or predictions of high or low values of each of the 10,001 trees grown. That is, each tree casts a vote of 0 or 1 ('no' or 'yes' to As >10 µg/L) for each cell based on the values of the predictor variables in that cell. Figures S2-S8 also provide more detailed views of the prediction map for each of the inhabited continents.

The importance of each of the 11 predictor variables in terms of mean decrease in accuracy and mean decrease in the Gini index is listed in Figure S1. Relative to the initial set of 52 variables, the values of these two statistics for most of the 11 final predictor variables appear to fall within a fairly narrow range, indicating comparable importance. Exceptions include fluvisols and soil pH, which have somewhat greater importance, and temperature, which according to both statistics is the least important of the 11 variables. Soil pH was also found to be an important predictor variable in arid, oxidizing environments in Pakistan(*29*). Although widespread arsenic dissolution occurs in Holocene fluvial sediments(*5-7, 9, 37*), this geological epoch has not been consistently mapped around the world. However, the global dataset of fluvisols provides a very suitable alternative(*29*), which may even be more appropriate as fluvisols by definition encompass recent fluvial sediments and not, for example, aeolian Holocene sediments that are generally not relevant for arsenic release. The generally high model importance of climate variables, as evidenced by them all being selected for the final model, highlights the strong control that climate has on arsenic release in aquifers. In particular, precipitation and evapotranspiration have a direct role in creating conditions conducive for arsenic release under reducing conditions (e.g. waterlogged soils) as well as high aridity associated with oxidizing high-pH conditions.

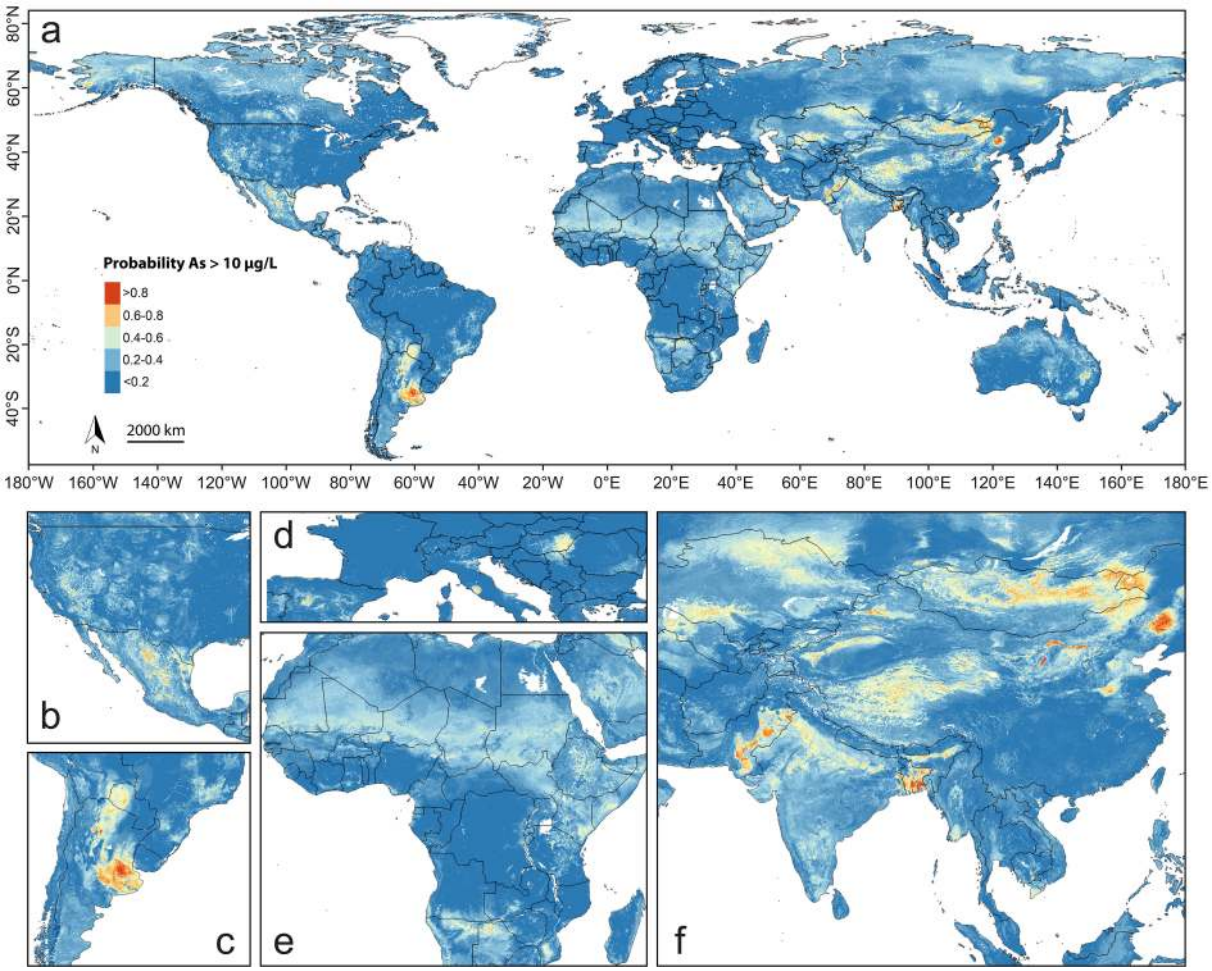


Figure 2. Modeled probability of arsenic concentration in groundwater exceeding 10 µg/L for the entire globe (a) along with zoomed-in sections of the main more densely populated affected area (b)-(f). The model is based on the arsenic data points in Figure 1 and the predictor variables in Table S2. Figures S2 to S8 provide more detailed views of the prediction map.

The performance of the random forest model on the test dataset (20% of the data, which was randomly selected while maintaining the relative distribution of high and low values) is summarized in the confusion matrix in Table 1. Despite a prevalence of high values (>10 µg/L) of only 22% in the dataset, the model performs well in predicting both high values (sensitivity: 0.79) and low values (specificity: 0.85) at a probability cutoff of 0.50. The average of these two figures, known as balanced accuracy, is correspondingly high at 0.82. Likewise, the model's AUC, which considers the full range of possible cutoffs, has a very high value of 0.89 with the test dataset (Table 1). For comparison, the AUC of a random forest using all 52 original predictor variables is also 0.89.

The model was also tested on a dataset of over 49,000 arsenic data points originating from known depths greater than 100 m (average 562 m, standard deviation 623 m). Although the model was not trained on any measurements from these depths and the fact that only surface

parameters were used as predictor variables, the model nevertheless performed quite well in predicting the arsenic concentrations of these deep groundwater sources, as evidenced by an AUC of 0.77.

Table 1. Confusion matrix and other statistics summarizing the results of applying the random forest model to the test dataset at a probability cutoff of 0.50.

	Measured	
	As ≤ 10 µg/L	As > 10 µg/L
Predicted As ≤ 10 µg/L	7710	555
Predicted As > 10 µg/L	1394	2037

Sensitivity: 0.79	Prevalence: 0.22
Specificity: 0.85	Balanced Accuracy: 0.82
Positive Predictive Value: 0.59	Cohen's kappa: 0.55
Negative Predictive Value: 0.93	AUC: 0.89

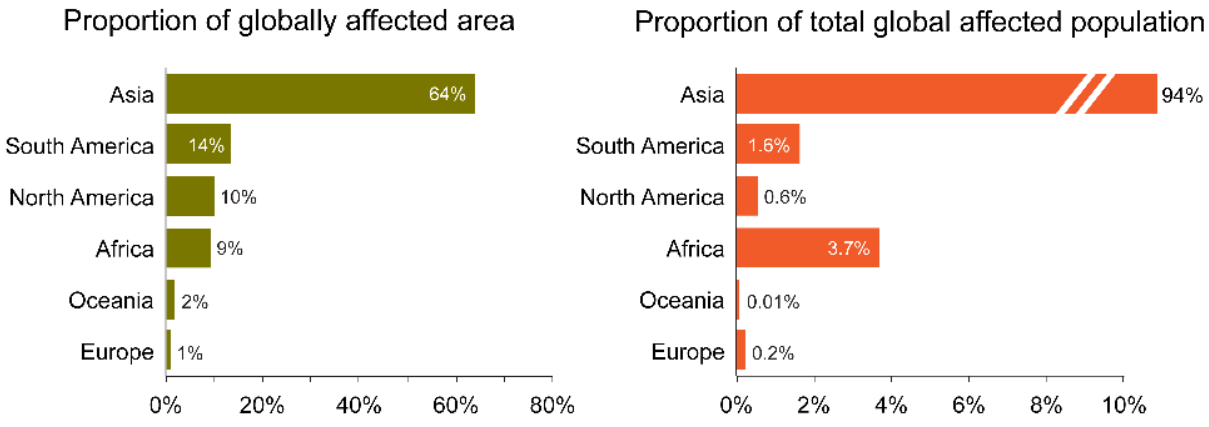


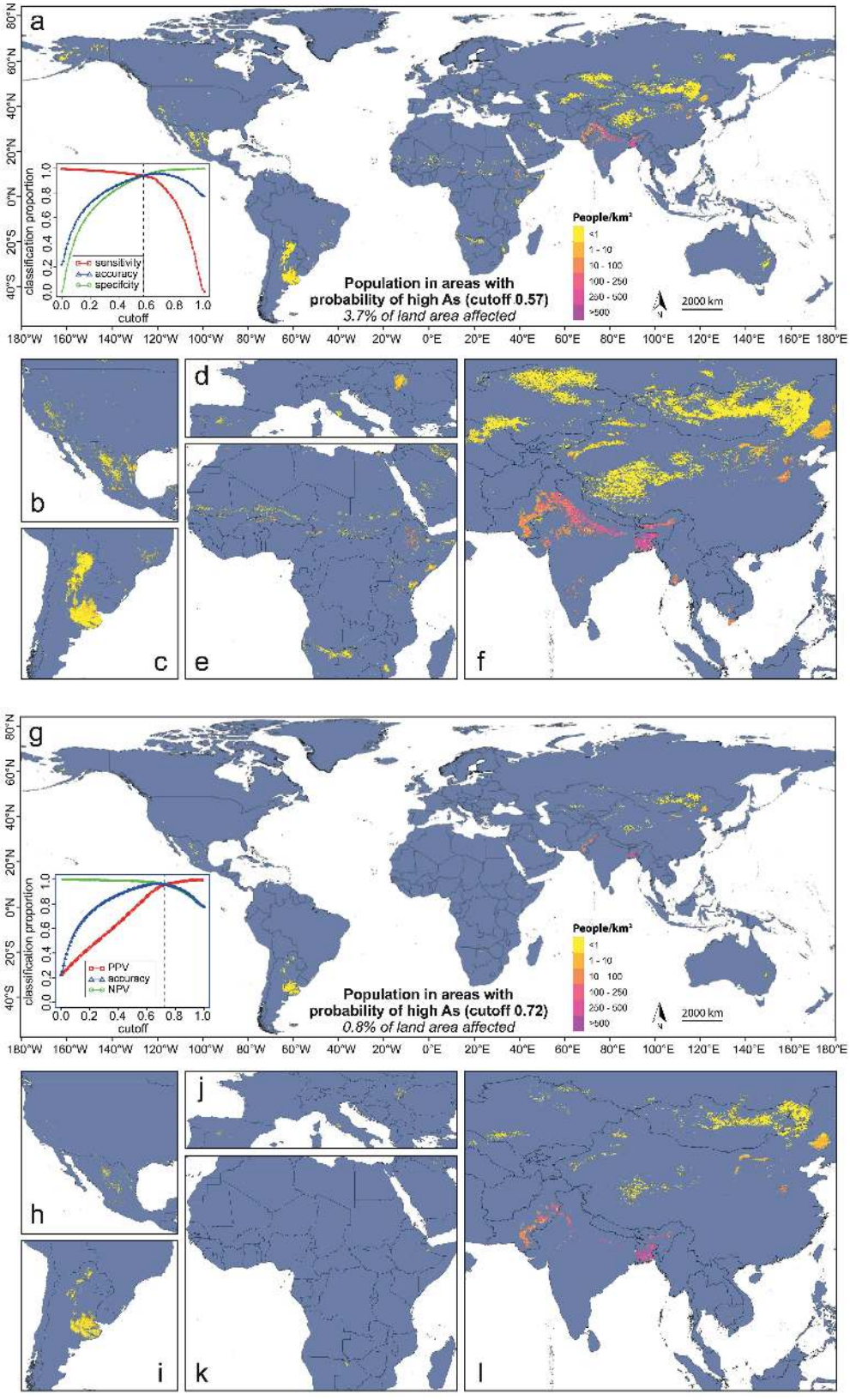
Figure 3. Proportions of land area and population potentially affected by arsenic concentrations in groundwater exceeding 10 µg/L by continent.

Regions and populations at risk

Areas predicted to have high arsenic concentrations in groundwater exist on all continents, with most being located in Central, South and Southeast Asia, parts of Africa and North and South America (Fig. 2 andFigs. S2-S8). Known areas of groundwater arsenic contamination are generally well captured by the global arsenic prediction map, e.g. parts of the western US, central Mexico, Argentina, the Pannonian Basin, Inner Mongolia, the Indus Valley, the Ganges-Brahmaputra Delta and the Mekong River and Red River Deltas. Areas of elevated arsenic hazard where little concentration data exist include parts of Central Asia, particularly Kazakhstan, Mongolia and Uzbekistan, the Sahel region and broad areas of the Arctic and sub-Arctic. Of these, the Central Asian hazard areas are better constrained as evidenced by higher probabilities.

Probability threshold values of 0.57 from the sensitivity-specificity comparison and 0.72 from the PPV-NPV comparison were found using the full dataset (combined training and test datasets) of arsenic concentrations. The proportions of high modeled arsenic hazard by continent associated with each of these probabilities are shown in Figure 3. Global maps of the potentially affected population in the risk areas as determined by these two thresholds are shown in Figure 4. As described in the Methods, these maps were then used to estimate the population potentially affected by drinking groundwater with arsenic concentrations exceeding 10 μg/L.

The resulting global arsenic risk assessment indicates that approximately 94-220 million people around the world (of which 85-90% are in South Asia) are potentially exposed to high concentrations of arsenic in groundwater from their domestic water supply (Tables S4 and S5). This range is consistent with the previous most comprehensive literature compilations, that is 140 million people(*41*) and 225 million people(*42*). Household groundwater-use statistics were not available for ~6-8 percent of the affected countries (depending on the cutoff), for which the less detailed statistics derived from the FAO AQUASTAT database were used instead (see Methods for details). To determine the amount of error that using these more general groundwater-use statistics might introduce to the overall population figures, the global potentially affected populations were recalculated with these countries' (those lacking household groundwater-use statistics) groundwater-use rates set to the extreme values of 0% and 100%. Since this applied to relatively few countries and As-affected areas, doing so affected the overall global population figures by an inconsequential amount (±0.1%), indicating that using the AQUASTAT groundwater-use rates, where necessary, is an acceptable approximation.



196

197 **Figure 4:** Population in risk areas potentially containing aquifers with arsenic concentrations
198 >10 µg/L using probability cutoffs of (a) 0.57, at which sensitivity and specificity are equal
199 (inset) as applied to the full (training and test) dataset, and (g) 0.72, at which PPV and NPV are
200 equal (inset) using the full dataset. The detailed areas of Fig. 2 are also repeated here for both
201 models (b)-(f) and (h)-(l).

This estimate of risk takes into account only the proportion of households utilizing unprocessed groundwater and assumes uniform rates throughout the urban and non-urban areas of each country. The uncertainties of these rates are unknown. The population in each cell was reduced by the uncertainty of the cell's prediction, which is justified based on the heterogeneity inherent in the accumulation of arsenic in an aquifer, which is generally at a much finer scale than that of the 1-km^2 resolution of the arsenic hazard map. Since the arsenic prediction for a cell represents the average outcome for that cell, we can take the modeled probability as a first-order approximation of the proportion of an aquifer in that cell containing high arsenic concentrations. Only cells exceeding the probability threshold (i.e. 0.57 or 0.72) were considered. The global estimate of 94-220 million people potentially affected by consuming arsenic-contaminated groundwater is broken down by continent and country in Tables S4 and S5, respectively, and represents the most accurate and consistent global estimate available.

DISCUSSION

The accuracy of the global groundwater arsenic prediction model presented here, as indicated, for example, with an AUC of 0.89 calculated with the test dataset, exceeds that found in previous arsenic prediction studies (Table S3). The dominance of climate and soil parameters in the final model is indicative of their direct influence or at least strong association with the processes of arsenic accumulation in groundwater.

With respect to previous arsenic prediction maps of global sedimentary basins(*40, 43*), the new model represents a significant advancement on a few different levels. First of all, the new model presented here provides predictions for all areas of the inhabited continents, whereas the previous first-generation statistical model covered only about half of the land areas. In addition, a ten-fold increase in measurement points has allowed arsenic concentrations to be incorporated from many more areas of the globe. The greatly expanded availability and quality of global predictor datasets over the past ten years has enabled new variables to be considered, such as soil type (e.g. fluvisols), as well as provided a 10 to 60-fold greater spatial resolution (i.e. 30 arc-seconds versus 5-30 arc-minutes). However, the presence of high arsenic in groundwater at a given location is of course predicated on the existence of an aquifer in the first place, which may not be so in the case of unfractured solid rock, steep terrain or very dry conditions. Models are only as good as the data on which they are based. As accurate as the new arsenic model is, it could be further improved as more arsenic data and more detailed predictor datasets come into existence.

Particularly in sedimentary aquifers, arsenic concentration is often highly dependent on depth, that is, to specific sedimentary sequences that differ in the concentration of arsenic in sediments as well as the geochemical conditions conducive to arsenic release. In order to better characterize this relationship in a given sedimentary basin, detailed depth information of groundwater samples would need to be incorporated in a separate basin-level study. Unfortunately, it is not feasible in a global-scale study to account for all of the diversity of the sedimentary basins of the world, especially since depth information of groundwater samples

is often not available. As such, we have relied on a statistical analysis of model performance against depth ranges of samples (where present) in order to determine model sensitivity to depth.

Our approach in the risk assessment of potentially affected population is relatively discerning and/or conservative. As such, the resulting population estimates may in some cases be lower than those found in earlier studies. One reason for this is that we employed country-specific statistics of rural and urban domestic groundwater usage, which allowed us to subtract the proportion of population utilizing surface water, tap water or other sources. This was not the case, for example, in a previous study of China that estimated 19.6 million people being affected in the country(*21*), whereas our estimate is considerably lower at 4.3-12.1 million. Furthermore, we consider only areas in which the probability of high arsenic exceeds the statistically determined cutoffs, i.e. 0.57 and 0.72. Taking the USA as an example, applying this criterion left only 0.2-2% of the area of the country over which to sum the potentially affected population (≤0.21 million, this study). In a previous arsenic risk assessment of the USA(*31*), the entire country was used to estimate affected population (2.1 million), that is, not only the high risk areas.

The actual proportion of groundwater usage varies spatially throughout a country, and so more detailed usage statistics beyond only urban versus rural would improve the accuracy of a risk assessment. In addition, more groundwater samples (ideally including depth information) from areas that currently have poor coverage would benefit future modeling efforts by allowing the model to be better adapted to those areas.

The presented arsenic probability maps should be used as a guide to further groundwater arsenic testing, for example in Central Asia, the Sahel and other regions of Africa. Only actual groundwater quality testing can definitively determine the suitability of groundwater with respect to arsenic, particularly due to small-scale (<1 km) aquifer heterogeneities that cannot be modeled with existing global datasets(*9, 44*). The hazard maps highlight areas at risk and provide a basis for targeted surveys, which continue to be important. The already large number of people potentially affected can be expected to increase as groundwater use expands with a growing population and increasing irrigation, especially in the light of water scarcity associated with warmer and drier conditions related to climate change. The maps can also help aid mitigation measures, such as awareness raising, coordination of government and financial support, health intervention programs, securing alternative drinking water resources and arsenic removal options tailored to the local groundwater conditions as well as social setting.

ACKNOWLEDGEMENTS

We thank our colleagues Anja Bretzler and Chris Zurbrügg (Eawag) and Andreas Steiner and Stephanie Piers de Raveschoot (SDC), and D.A. Polya and R. Wu (University of Manchester) for their support as well as the many providers of data, which were an essential component of

this work. **Funding:** We thank the Swiss Agency for Development and Cooperation (projects no. 7F-09010.01.01 and 7F-09963.01.01) for their long-term support and co-funding of this study, as well as a University of Manchester EPSRC IAA Impact Support Fund Award. **Author contributions:** J.P.: Methodology, Modeling, Writing- Original draft preparation, M.B.: Supervision, Writing- Reviewing and Editing. **Competing interests:** Authors declare no competing interests. **Data and materials availability:** The modeling data, code and raster output maps are available at ERIC/open(*45*). Arsenic concentration data points and hazard and risk maps are also available for viewing on the GIS-based Groundwater Assessment Platform (GAP), www.gapmaps.org.

Supplementary Materials

Methods
Supplementary Tables 1-6
Supplementary Figures 1-11
References 46-125

References and notes

1. A. H. Smith, E. O. Lingas, M. Rahman, Contamination of drinking-water by arsenic in Bangladesh: a public health emergency. *Bulletin of the World Health Organization* **78**, 1093-1103 (2000).
2. M. F. Hughes, Arsenic toxicity and potential mechanisms of action. *Toxicology letters* **133**, 1-16 (2002).
3. D. Mondal *et al.*, Comparison of drinking water, raw rice and cooking of rice as arsenic exposure routes in three contrasting areas of West Bengal, India. *Environmental geochemistry and health* **32**, 463-477 (2010).
4. WHO, Guidelines for drinking-water quality. *WHO chronicle* **38**, 104-108 (2011).
5. P. Smedley, D. Kinniburgh, A review of the source, behaviour and distribution of arsenic in natural waters. *Applied geochemistry* **17**, 517-568 (2002).
6. R. Nickson *et al.*, Arsenic poisoning of Bangladesh groundwater. *Nature* **395**, 338-338 (1998).
7. J. McArthur, P. Ravenscroft, S. Safiulla, M. Thirlwall, Arsenic in groundwater: testing pollution mechanisms for sedimentary aquifers in Bangladesh. *Water Resources Research* **37**, 109-117 (2001).
8. M. Berg *et al.*, Hydrological and sedimentary controls leading to arsenic contamination of groundwater in the Hanoi area, Vietnam: the impact of iron-arsenic ratios, peat, river bank deposits, and excessive groundwater abstraction. *Chemical Geology* **249**, 91-112 (2008).
9. S. Fendorf, H. A. Michael, A. van Geen, Spatial and temporal variations of groundwater arsenic in South and Southeast Asia. *Science* **328**, 1123-1127 (2010).
10. M. I. Litter *et al.*, Arsenic in Argentina: Occurrence, human health, legislation and determination. *Science of The Total Environment*, (2019).
11. Y. Zheng, S. V. Flanagan, The case for universal screening of private well water quality in the US and testing requirements to achieve it: evidence from arsenic. *Environmental health perspectives* **125**, 085002 (2017).
12. P. Bhattacharya, D. Chatterjee, G. Jacks, Occurrence of Arsenic-contaminatedGroundwater in Alluvial Aquifers from Delta Plains, Eastern India: Options for Safe Drinking Water Supply. *International Journal of Water Resources Development* **13**, 79-92 (1997).

13. A. Van Geen *et al.*, Spatial variability of arsenic in 6000 tube wells in a 25 km2 area of Bangladesh. *Water Resources Research* **39**, (2003).
14. H. B. Nicolli, J. M. Suriano, M. A. G. Peral, L. H. Ferpozzi, O. A. Baleani, Groundwater contamination with arsenic and other trace elements in an area of the Pampa, Province of Córdoba, Argentina. *Environmental Geology and Water Sciences* **14**, 3-16 (1989).
15. P. Smedley, H. Nicolli, D. Macdonald, A. Barros, J. Tullio, Hydrogeochemistry of arsenic and other inorganic constituents in groundwaters from La Pampa, Argentina. *Applied Geochemistry* **17**, 259-284 (2002).
16. M. Blarasin, A. Cabrera, E. Matteoda, paper presented at the XXXIII IAH -. 7º ALHSUD Congress, Zacatecas, Mexico, 11 - 15 October 2004 2004.
17. M. Auge, G. E. Viale, L. Sierra, in *VIII Congreso Argentino de Hidrogeología: Aguas subterráneas recurso estratégico*. (2013), vol. 2, pp. 58-63.
18. M. Berg *et al.*, Arsenic contamination of groundwater and drinking water in Vietnam: a human health threat. *Environmental Science & Technology* **35**, 2621-2626 (2001).
19. J. Buschmann, M. Berg, C. Stengel, M. L. Sampson, Arsenic and manganese contamination of drinking water resources in Cambodia: coincidence of risk areas with low relief topography. *Environmental science & technology* **41**, 2146-2152 (2007).
20. P. Smedley, M. Zhang, G. Zhang, Z. Luo, Mobilisation of arsenic and other trace elements in fluviolacustrine aquifers of the Huhhot Basin, Inner Mongolia. *Applied Geochemistry* **18**, 1453-1477 (2003).
21. L. Rodríguez-Lado *et al.*, Groundwater arsenic contamination throughout China. *Science* **341**, 866-868 (2013).
22. Y. Zhou *et al.*, Distribution of groundwater arsenic in Xinjiang, PR China. *Applied geochemistry* **77**, 116-125 (2017).
23. D. Chatterjee, R. Roy, B. Basu, Riddle of arsenic in groundwater of Bengal Delta Plain—role of non-inland source and redox traps. *Environmental Geology* **49**, 188-206 (2005).
24. B. Nath, D. Stüben, S. B. Mallik, D. Chatterjee, L. Charlet, Mobility of arsenic in West Bengal aquifers conducting low and high groundwater arsenic. Part I: Comparative hydrochemical and hydrogeological characteristics. *Applied Geochemistry* **23**, 977-995 (2008).
25. B. A. Shah, Arsenic-contaminated groundwater in Holocene sediments from parts of middle Ganga plain, Uttar Pradesh, India. *Current Science(Bangalore)* **98**, 1359-1365 (2010).
26. B. Planer-Friedrich, *Hydrogeological and hydrochemical investigations in the Rioverde basin, Mexico*. (Verlag nicht ermittelbar, 2000).
27. M. T. Alarcón-Herrera *et al.*, Co-occurrence of arsenic and fluoride in groundwater of semi-arid regions in Latin America: Genesis, mobility and remediation. *Journal of Hazardous Materials* **262**, 960-969 (2013).
28. R. Nickson, J. McArthur, B. Shrestha, T. Kyaw-Myint, D. Lowry, Arsenic and other drinking water quality issues, Muzaffargarh District, Pakistan. *Applied Geochemistry* **20**, 55-68 (2005).
29. J. E. Podgorski *et al.*, Extensive arsenic contamination in high-pH unconfined aquifers in the Indus Valley. *Science Advances* **3**, (2017).
30. J. D. Ayotte, M. G. Nielsen, G. R. Robinson Jr, R. B. Moore, Relation of arsenic, iron, and manganese in ground water to aquifer type, bedrock lithogeochemistry, and land use in the New England Coastal Basins. *Water Resources Investigations Report* **99**, 4162 (1999).
31. J. D. Ayotte, L. Medalie, S. L. Qi, L. C. Backer, B. T. Nolan, Estimating the high-arsenic domestic-well population in the conterminous United States. *Environ. Sci. Technol.* **51**, 12443-12454 (2017).
32. M. Berg *et al.*, Magnitude of arsenic pollution in the Mekong and Red River Deltas—Cambodia and Vietnam. *Science of the Total Environment* **372**, 413-425 (2007).
33. L. H. Winkel *et al.*, Arsenic pollution of groundwater in Vietnam exacerbated by deep aquifer exploitation for more than a century. *Proceedings of the National Academy of Sciences* **108**, 1246-1251 (2011).
34. A. Bretzler *et al.*, Groundwater arsenic contamination in Burkina Faso, West Africa: Predicting and verifying regions at risk. *Science of the Total Environment* **584**, 958-970 (2017).

35. Q. Zhang *et al.*, Coupling predicted model of arsenic in groundwater with endemic arsenism occurrence in Shanxi Province, Northern China. *Journal of hazardous materials* **262**, 1147-1153 (2013).
36. S. Bindal, C. K. Singh, Predicting groundwater arsenic contamination: regions at risk in highest populated state of India. *Water Res* **159**, 65-76 (2019).
37. L. Winkel, M. Berg, M. Amini, S. J. Hug, C. A. Johnson, Predicting groundwater arsenic contamination in Southeast Asia from surface parameters. *Nat. Geosci.* **1**, 536-542 (2008).
38. Q. Yang, H. B. Jung, R. G. Marvinney, C. W. Culbertson, Y. Zheng, Can arsenic occurrence rates in bedrock aquifers be predicted? *Environmental science & technology* **46**, 2080-2087 (2012).
39. N. Yang, L. H. Winkel, K. H. Johannesson, Predicting geogenic arsenic contamination in shallow groundwater of South Louisiana, United States. *Environmental science & technology* **48**, 5660-5666 (2014).
40. M. Amini *et al.*, Statistical modeling of global geogenic arsenic contamination in groundwater. *Environmental science & technology* **42**, 3669-3675 (2008).
41. P. Ravenscroft, H. Brammer, K. Richards, *Arsenic pollution: a global synthesis*. (John Wiley & Sons, 2009), vol. 28.
42. S. Murcott, *Arsenic contamination in the world*. (IWA publishing, 2012).
43. P. Ravenscroft, Predicting the global extent of arsenic pollution of groundwater and its potential impact on human health. *Unpublished report prepared for UNICEF, December*, (2007).
44. Y. Zheng, Lessons learned from arsenic mitigation among private well households. *Current environmental health reports* **4**, 373-382 (2017).
45. J. Podgorski, M. Berg, Podgorski_and_Berg_2020. *ERIC/open* (2020); doi.org/10.25678/0001ZT.
46. L. Breiman, Random forests. *Machine learning* **45**, 5-32 (2001).
47. R Core Team*,* R Foundation for Statistical Computing, Ed. (Vienna, Austria, 2014).
48. T. T. Hastie, Robert; Friedman, Jerome, *The Elements of Statistical Learning (2nd ed.)*. (Springer, 2008).
49. T. K. Ho, in *Document analysis and recognition, 1995., proceedings of the third international conference on*. (IEEE, 1995), vol. 1, pp. 278-282.
50. M. L. McHugh, Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica* **22**, 276-282 (2012).
51. T. Fawcett, An introduction to ROC analysis. *Pattern recognition letters* **27**, 861-874 (2006).
52. J. Gao*,* NASA Socioeconomic Data and Applications Center (SEDAC), Ed. (Palisades, NY, 2019).
53. JMP. (WHO/UNICEF Joint Monitoring Program (JMP), 2019), vol. 2019.
54. M. A. Friedl *et al.*, MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote sensing of Environment* **114**, 168-182 (2010).
55. FAO*,* Food and Agriculture Organization of the United Nations (FAO), Ed. (2016).
56. R. E. Broshears, M. A. Akbari, M. P. Chornack, D. K. Mueller, B. C. Ruddy, "Inventory of ground-water resources in the Kabul Basin, Afghanistan," (U. S. Geological Survey, 2005).
57. UNHCR. (UNHCR, 2019), vol. 2019.
58. M. E. Zabala, M. Manzano, L. Vives, Assessment of processes controlling the regional distribution of fluoride and arsenic in groundwater of the Pampeano Aquifer in the Del Azul Creek basin (Argentina). *Journal of hydrology* **541**, 1067-1087 (2016).
59. M. E. Morgada, M. Mateu, J. Bundschuh, M. I. Litter, Arsenic in the Iberoamerican region. The IBEROARSEN Network and a possible economic solution for arsenic removal in isolated rural zones. *e-Terra* **5**, 1-11 (2008).
60. K. Ivkovic, K. Watkins, R. Cresswell, J. Bauld, A groundwater quality assessment of the fractured rock aquifers of the Piccadilly Valley, South Australia. (1998).
61. J. Fitzgerald *et al.*, Groundwater quality and environmental health implications. *Anangu Pitjantjara Lands, South Australia, A Report from Bureau of Rural Sciences*, 1-30 (1999).
62. S. Clohessy, "Perth Shallow Groundwater Systems Investigation: Lake Gwelup," *Hydrogeological record series* (Department of Water, Perth, 2012).

63. R. M. Larsen, *A Groundwater Quality Assessment of the Jandakot Mound, Swan Coastal Plain, Western Australia*. (Australian Geological Survey Organisation, 1998).
64. E. E. A. EEA. (2019).
65. D. Kinniburgh, P. Smedley, Arsenic contamination of groundwater in Bangladesh. (2001).
66. *DWA Maun Groundwater Development Project: Phase 2, Resources assessment and wellfield development: Final report,* (2004).
67. CPRM*,* Geological Survey of Brazil, Ed. (2017), vol. 2017.
68. Ministry of Rural Development of Cambodia. (2015).
69. E. K. Read *et al.*, Water quality data for national-scale aquatic research: The Water Quality Portal. *Water Resources Research* **53**, 1735-1745 (2017).
70. D. R. Boyle, W. A. Spirito, S. W. Adcock, "Groundwater hydrogeochemical survey of central New Brunswick," (1996).
71. C. Reimann, K. Bjorvatn, R. Tekle-Haimanot, Z. Melako, U. Siewers, Drinking water quality, Rift Valley, Ethiopia. *Norges geologiske undersøkelse, Report* **2002**, 132 (2002).
72. A. Bretzler *et al.*, Groundwater origin and flow dynamics in active rift systems–A multi-isotope approach in the Main Ethiopian Rift. *Journal of hydrology* **402**, 274-289 (2011).
73. T. Rango, G. Bianchini, L. Beccaluva, R. Tassinari, Geochemistry and water quality assessment of central Main Ethiopian Rift natural waters with emphasis on source and occurrence of fluoride and arsenic. *Journal of African Earth Sciences* **57**, 479-491 (2010).
74. B. Kortatsi *et al.*, Reconnaissance survey of arsenic concentration in ground-water in south-eastern Ghana. *West African Journal of Applied Ecology* **13**, 16-26 (2008).
75. P. L. Smedley, Arsenic in rural groundwater in Ghana: part special issue: hydrogeochemical studies in sub-Saharan Africa. *Journal of African Earth Sciences* **22**, 459-470 (1996).
76. I. A. Katsoyiannis, S. J. Hug, A. Ammann, A. Zikoudi, C. Hatziliontos, Arsenic speciation and uranium concentrations in drinking water supply wells in Northern Greece: correlations with redox indicative parameters and implications for groundwater treatment. *Science of the Total Environment* **383**, 128-140 (2007).
77. H. A. Rowland *et al.*, Geochemistry and arsenic behaviour in groundwater resources of the Pannonian Basin (Hungary and Romania). *Applied Geochemistry* **26**, 1-17 (2011).
78. S. Chandra, S. Ahmed, E. Nagaiah, S. K. Singh, P. Chandra, Geophysical exploration for lithological control of arsenic contamination in groundwater in Middle Ganga Plains, India. *Physics and Chemistry of the Earth, Parts A/B/C* **36**, 1353-1362 (2011).
79. T. Ghosh, R. Kanchan, Geoenvironmental appraisal of groundwater quality in Bengal alluvial tract, India: a geochemical and statistical approach. *Environmental earth sciences* **72**, 2475-2488 (2014).
80. A. Mukherjee *et al.*, Controls on high and low groundwater arsenic on the opposite banks of the lower reaches of River Ganges, Bengal basin, India. *Science of the Total Environment* **645**, 1371-1387 (2018).
81. J. McArthur *et al.*, How paleosols influence groundwater flow and arsenic pollution: a model from the Bengal Basin and its worldwide implication. *Water Resources Research* **44**, (2008).
82. A. Mukherjee, A. E. Fryar, H. D. Rowe, Regional-scale stable isotopic signatures of recharge and deep groundwater in the arsenic affected areas of West Bengal, India. *Journal of Hydrology* **334**, 151-161 (2007).
83. V. S. Chauhan, R. Nickson, D. Chauhan, L. Iyengar, N. Sankararamakrishnan, Ground water geochemistry of Ballia district, Uttar Pradesh, India and mechanism of arsenic release. *Chemosphere* **75**, 83-91 (2009).
84. D. Saha, S. Sahu, A decade of investigations on groundwater arsenic contamination in Middle Ganga Plain, India. *Environmental geochemistry and health* **38**, 315-337 (2016).
85. D. P. Shukla, C. Dubey, N. P. Singh, M. Tajbakhsh, M. Chaudhry, Sources and controls of Arsenic contamination in groundwater of Rajnandgaon and Kanker District, Chattisgarh Central India. *Journal of Hydrology* **395**, 49-66 (2010).

86. J. P. Maity *et al.*, Arsenic-enriched groundwaters of India, Bangladesh and Taiwan—Comparison of hydrochemical characteristics and mobility constraints. *Journal of Environmental Science and Health, Part A* **46**, 1163-1176 (2011).
87. S. Kar *et al.*, Arsenic-enriched aquifers: occurrences and mobilization of arsenic in groundwater of Ganges Delta Plain, Barasat, West Bengal, India. *Applied Geochemistry* **25**, 1805-1814 (2010).
88. S. Hazarika, B. Bhuyan, Fluoride, arsenic and iron content of groundwater around six selected tea gardens of Lakhimpur District, Assam, India. *Arch Appl Sci Res* **5**, 57-61 (2013).
89. B. Nath *et al.*, Hydrochemistry of arsenic-enriched aquifer from rural West Bengal, India: a study of the arsenic exposure and mitigation option. *Water, air, and soil pollution* **190**, 95-113 (2008).
90. R. A. Olea, N. J. Raju, J. J. Egozcue, V. Pawlowsky-Glahn, S. Singh, Advancements in hydrochemistry mapping: methods and application to groundwater arsenic and iron concentrations in Varanasi, Uttar Pradesh, India. *Stochastic environmental research and risk assessment* **32**, 241-259 (2018).
91. M. Kumar, A. Ramanathan, M. M. Rahman, R. Naidu, Concentrations of inorganic arsenic in groundwater, agricultural soils and subsurface sediments from the middle Gangetic plain of Bihar, India. *Science of the Total Environment* **573**, 1103-1114 (2016).
92. S. Chidambaram *et al.*, A study on the arsenic concentration in groundwater of a coastal aquifer in south-east India: an integrated approach. *Environment, Development and Sustainability* **19**, 1015-1040 (2017).
93. S. Ghosh, P. Sar, Identification and characterization of metabolic properties of bacterial populations recovered from arsenic contaminated ground water of North East India (Assam). *Water research* **47**, 6992-7005 (2013).
94. S. Sharma, J. Kaur, A. K. Nagpal, I. Kaur, Quantitative assessment of possible human health risk associated with consumption of arsenic contaminated groundwater and wheat grains from Ropar Wetand and its environs. *Environmental monitoring and assessment* **188**, 506 (2016).
95. B. A. Shah, Role of Quaternary stratigraphy on arsenic-contaminated groundwater from parts of Barak Valley, Assam, North–East India. *Environmental earth sciences* **66**, 2491-2501 (2012).
96. B. A. Shah, Role of Quaternary stratigraphy on arsenic-contaminated groundwater from parts of Middle Ganga Plain, UP–Bihar, India. *Environmental geology* **53**, 1553-1561 (2008).
97. B. A. Shah, Status of groundwater arsenic pollution of Mirzapur district in Holocene aquifers from parts of the Middle Ganga Plain, India. *Environmental earth sciences* **73**, 1505-1514 (2015).
98. L. Sailo, C. Mahanta, Arsenic mobilization in the Brahmaputra plains of Assam: groundwater and sedimentary controls. *Environmental monitoring and assessment* **186**, 6805-6820 (2014).
99. D. Paul, S. K. Kazy, A. K. Gupta, T. Pal, P. Sar, Diversity, metabolic properties and arsenic mobilization potential of indigenous bacteria in arsenic contaminated groundwater of West Bengal, India. *PloS one* **10**, e0118735 (2015).
100. Catholic Relief Services, Meulaboh Water Quality Laboratory in Aceh, Ed. (2007).
101. UNEP, "Water Quality, 2005 State of the UNEP GEMS/Water Global Network and Annual Report," (2005).
102. M. Pritchard, T. Mkandawire, J. O'neill, Assessment of groundwater quality in shallow wells within the southern districts of Malawi. *Physics and Chemistry of the Earth, Parts A/B/C* **33**, 812-823 (2008).
103. INCA, *Arsénico y fluoruro en agua: riesgos y perspectivas desde la sociedad civil y la academia en México*. (Mexico, 2018).
104. A. Van Geen *et al.*, Confirmation of elevated arsenic levels in groundwater of Myanmar. *Science of the Total Environment* **478**, 21-24 (2014).
105. B. R. Shrestha, J. W. Whitney, K. B. Shrestha, "The State of Arsenic in Nepal-2003," (Kathmandu, Nepal, 2004).

106. B. Frengstad, A. K. M. Skrede, D. Banks, J. R. Krog, U. Siewers, The chemistry of Norwegian groundwaters: III. The distribution of trace elements in 476 crystalline bedrock groundwaters, as analysed by ICP-MS techniques. *Science of the Total environment* **246**, 21-40 (2000).
107. P. de Caritat, S. Danilova, C. Reimann, G. Storrø, Groundwater composition near the nickel—copper smelting industry on the Kola Peninsula, central Barents Region (NW Russia and NE Norway). *Journal of Hydrology* **208**, 92-107 (1998).
108. C. M. de Meyer *et al.*, Arsenic, manganese and aluminum contamination in groundwater resources of Western Amazonia (Peru). *Science of the Total Environment* **607**, 1437-1450 (2017).
109. L. McCaffrey, J. Willis, *Distribution of fluoride-rich groundwater in the eastern and Mogwase regions of the Northern and North-West Provinces*. (Water Research Commission Pretoria, 2001).
110. Geological Survey of Sweden. (2007).
111. M. Haldimann, E. Pfammatter, P.-M. Venetz, P. Studer, V. Dudler, Occurrence of arsenic in drinking water of the canton of Valais. Part I: Overview of arsenic concentration and geographic distribution. *Mitteilungen aus Lebensmitteluntersuchung und Hygiene* **96**, 89-105 (2005).
112. P. Smedley *et al.*, Fluoride in groundwater from high-fluoride areas of Ghana and Tanzania. (2002).
113. J. Buschmann *et al.*, Contamination of drinking water resources in the Mekong delta floodplains: Arsenic and other trace metals pose serious health risks to population. *Environment International* **34**, 756-764 (2008).
114. A. Trabucco, R. Zomer, Global soil water balance geospatial database. *CGIAR Consortium for Spatial Information, Published online, available from the CGIAR-CSI GeoPortal at: http://www.cgiar-csi.org (last access: January 2013)*, (2010).
115. A. Trabucco, R. J. Zomer, Global aridity index (global-aridity) and global potential evapo-transpiration (global-PET) geospatial database. *CGIAR Consortium for Spatial Information*, (2009).
116. R. J. Hijmans, S. E. Cameron, J. L. Parra, P. G. Jones, A. Jarvis, Very high resolution interpolated climate surfaces for global land areas. *International journal of climatology* **25**, 1965-1978 (2005).
117. S. E. Fick, R. J. Hijmans, WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, (2017).
118. J. Hartmann, N. Moosdorf, The new global lithological map database GLiM: A representation of rock properties at the Earth surface. *Geochemistry, Geophysics, Geosystems* **13**, (2012).
119. C. E. R. S. C. U.S. Geological Survey - Energy Resources Program. (Department of the Interior/USGS).
120. T. Hengl *et al.*, SoilGrids250m: Global gridded soil information based on machine learning. *PLoS one* **12**, e0169748 (2017).
121. C. W. Ross *et al.*, HYSOGs250m, global gridded hydrologic soil groups for curve-number-based runoff modeling. *Scientific data* **5**, 180091 (2018).
122. J. Pelletier *et al.*, Global 1-km gridded thickness of soil, regolith, and sedimentary deposit layers. *ORNL DAAC*, (2016).
123. USGS. (U.S. Geological Survey, EROS Data Cent. Sioux Falls, SD, 1996).
124. T. Hengl. (Zenodo, 2018).
125. Y. Fan, H. Li, G. Miguez-Macho, Global patterns of groundwater table depth. *Science* **339**, 940-943 (2013).